

In: **Computation, Causation, and Discovery**. Eds. P Glymour and G. Cooper.
Menlo Park, CA, Cambridge, MA: AAAI Press / The MIT Press. 1999. pp. 333-342.

CHAPTER TEN

Rejoinder to Glymour and Spirtes

James M. Robins and Larry Wasserman

1. Introduction

The opening sentence to the abstract of our paper is the statement that Spirtes, Glymour, and Scheines and Pearl and Verma make the startling claim that it is possible to infer causal relationships between two variables X and Y from associations found in observational data without substantive subject-matter-specific background knowledge. Such a claim seemed, on its face, specious to us and to other epidemiologists, statisticians, and social scientists we polled.

What was fascinating to us was (1) that Spirtes, Glymour, and Scheines and Pearl and Verma proved, given their causal framework (i.e., in the absence of selection bias, the observed distributions of a set of variables is the marginal of a distribution faithful to some causally sufficient DAG) that, without substantive subject matter knowledge, there exist methods for discovering causal structure from observational data which are asymptotically correct, and (2) that we did not find their causal framework or their faithfulness assumption to be unreasonable.

However, we then recognized that implicit in their asymptotics was the assumption that the prior probability of there being no unmeasured confounders was positive and not small in relation to sample size. We proved that without this additional assumption, Spirtes, Glymour, and Scheines's and Pearl and Verma's methods cannot be reliably used to infer causal structure from observational data. Having read Glymour, Spirtes, and Richardson's reply, we continue to believe that essentially any pair of variables is dependent due to unmeasured confounders. Indeed, as we argued in the chapter, we believe the following proposition (which we also believe is well accepted

by practicing epidemiologists, statisticians, and social scientists): due to residual unmeasured confounding, essentially any two causally unconnected variables will be found to have a highly statistically significant association in most very large observational studies. Glymour, Spirtes, and Richardson argue in their section 3 that we offered no convincing evidence that practicing epidemiologists endorsed this proposition. However, we will show below that the Glymour, Spirtes, and Richardson's argument is flawed.

In section 3 of their reply, Glymour, Spirtes, and Richardson concede one of our main points: in observational studies, small causal effects cannot be reliably ruled in nor ruled out, no matter how large the study size. In the final sections of our rejoinder, we give our view on the possible implications of this concession as to the reliability of causal searches based on the Tetrad program.

2. The Loss Function

We now turn to a detailed consideration of Glymour, Spirtes, and Richardson's argument in their reply. At the end of their section 1, Glymour, Spirtes, and Richardson raised two questions: (1) how plausible our priors are upon reflection, and (2) whether such priors and loss functions are actually used in practice. We agree these are critical questions to which we now give answers.

To begin, our decision to focus on the zero-one loss function (i.e., whether a variable was or was not the cause of another variable) rather than the magnitude of any causal effect was not because we think the zero-one loss function to be the most relevant—we don't; rather, it is because Spirtes, Glymour, and Scheines and Pearl and Verma themselves emphasize the zero-one loss functions, it being implicit whenever they say they can reliably determine causal structure. Indeed, nowhere in Spirtes, Glymour, and Scheines or Pearl and Verma is there, as far as we know, any extended discussion of the fact that we emphasized in our chapter: small causal effects cannot be reliably ruled out or ruled in based on observational data, no matter how large the sample size.

The impact of this point is ameliorated to some extent by the fact that, for policy purposes, the loss associated with misidentifying small causal effects is usually less than that associated with misidentifying large causal effects. However, the scale on which the loss associated with misidentifying a causal effect is measured for policy purposes may differ from the scale on which identification of causal effects can be reliably determined, as the following discussion illustrates.

In epidemiology, it is commonly argued that one of the main criteria for

determining the likelihood of an association as a causal association is its magnitude (Hill 1951). However, as stressed by Robins et al. (1985), for a dichotomous outcome, this magnitude is to be measured on a ratio (i.e., relative) scale, not on a difference (i.e., absolute) scale.

As an example, given sufficiently large studies so that sampling variability does not dominate uncertainty, a tenfold increase in the risk of soft tissue sarcoma associated with an occupational exposure A is more likely to be causal than a twenty percent increase in the risk of coronary heart disease associated with an exposure B , all else being equal. This statement reflects both the fact that it is harder to imagine that bias due to unmeasured common causes, measurement error, and selection bias could easily explain a risk ratio of eleven than a risk ratio of 1.2.

Yet the public health benefits of preventing exposure to a causal agent are properly measured on a difference rather than ratio scale. Specifically, because the background rate of coronary heart disease is so many times greater than the background rate of soft tissue sarcoma, the number of excess deaths associated with a true twenty percent increase in coronary heart disease mortality due to agent B would exceed by seven-fold the excess deaths associated with a true tenfold increase in the risk of soft tissue sarcoma. It follows that any method that will allow us to reliably determine from observational data whether an increase in coronary heart disease of twenty percent was likely to be causal would be invaluable. The Tetrad project, as described in Spirtes, Glymour, and Scheines (1993), would seem to offer such promise. However, for the reasons we discuss in our chapter, we are skeptical. We believe that no purely statistical methods can ever help us to determine whether small relative increases in risk are causal.

3. Plausibility of Priors

We turn now to consider the plausibility of our priors. In section 2, Glymour, Spirtes, and Richardson criticize the prior we worked with in section 2. As we stress in the chapter, this prior is not meant to represent the actual subjective beliefs of anyone. It was chosen for only one purpose: to make the formal Bayesian calculations we propose transparent and explicit. Glymour, Spirtes, and Richardson state that their comments also are relevant to the realistic priors of our section 6. We disagree with this claim. We regard Glymour, Spirtes, and Richardson's section 2 largely as an attack on our highly unrealistic prior of section 2, rather than on the realistic priors discussed in our section 6. It will be helpful here if we here restate our main conclusions concerning the realistic priors of our section 6.

Let n be the sample size and $P(N)$ be the prior probability of the event N

that there are no unmeasured common causes of X and Y where X is known to be temporally prior to Y . Then under the minimal assumptions on our prior described in our section 6, our theorem 3 can be restated as follows.

If the probability of there being no unmeasured common causes is small relative to sample size [specifically, $P(N) = o(n^{-1/2})$], then, even when X and Y are exactly uncorrelated in the data, the posterior odds that X does not cause Y converges as $n \rightarrow \infty$ to a number ψ that depends on the prior odds that X does not cause Y and ψ is bounded away from zero in infinity. As a consequence, the causal relationship that X does not cause Y cannot be determined by the data alone, no matter how large the sample size, even assuming faithfulness.

Thus, in view of our theorem 3, and as Glymour, Spirtes, and Richardson recognize in their section 3, the important question is not how reasonable is our simplistic prior of section 2 but rather (1) do practicing epidemiologists, statisticians, and social scientists have prior beliefs satisfying $P(N) = o(n^{-1/2})$, and (2) should they still hold these beliefs after reading the discussion in Glymour, Spirtes, and Richardson?

To determine the beliefs of practicing professionals, we asked six epidemiologists at Harvard and UCLA the following question. Consider twenty large epidemiologic studies (say, with sample size 240,000 or greater). Consider any dichotomous exposure X and any dichotomous outcome Y which you believe are not causally connected and that have been measured in each of the twenty studies. Suppose further the prevalence of each variable is approximately 50 percent. What is your best guess as to the fraction of the twenty studies in which the empirical odds ratio for the two variables OR_{XY} will lie outside of the interval (.98, 1.02)?

Of the epidemiologists queried, the mean was 90 percent with a minimum of 75 percent. Since the empirical odds ratio lying outside of this interval implies that the p -value for the usual χ^2 test of independence is less than .02, we conclude that these epidemiologists believe that with high probability, any two variables will be correlated in large observational studies.

Furthermore, the mean fell only to 85 percent when we further told the epidemiologists that known confounding factors had been adjusted for. In addition, all of the epidemiologists agreed with the statement that, even when the odds ratio was contained in the interval (.98, 1.02), this fact should not be taken as strong evidence that X and Y are uncorrelated, because there may be inadequate power to detect small but nonzero correlations at the $p < .02$ level.

When questioned further as to whether such associations were due to confounding, selection bias, or both, most epidemiologists chose the combination of the two. However, it is easy to show that the arguments in section 6 of our chapter go through unchanged if we consider $P(N)$ as the prior probability of the event that there is neither confounding by unmeasured common causes nor selection bias (i.e., due to conditioning on a common effect of X

and Y). Henceforth, as is common in epidemiology, we shall say that such noncausal associations between X and Y are due to unmeasured confounding (regardless of whether this confounding is due to an unmeasured common cause or to conditioning on a common effect).

We conclude that the above small poll is consistent with our statement that practicing epidemiologists believe the proposition that, due to unmeasured confounding, with probability nearly 1, two causally unconnected variables X and Y will be associated in the population.

How is it that Glymour, Spirtes, and Richardson argue that we offer no convincing evidence that epidemiologists believe this proposition? We believe their argument derives from a misreading of our chapter. (The way the chapter was written partly invited such a misreading.) In our section 7 on prior beliefs, it is in the second paragraph that we discuss where beliefs in the above proposition derive from. Unfortunately, Glymour, Spirtes, and Richardson attack the arguments we make in the first paragraph of section 7, which were meant to be only subsidiary to the central arguments given in the second paragraph.

Specifically, in the first paragraph of section 7, we state that subjective beliefs of epidemiologists hold that the prior probability of no unmeasured confounders is extremely small, if not zero, for otherwise, if the sample size n were large, then when X and Y are uncorrelated in the data, we could reliably conclude that X does not cause Y and thus rule out the small causal effects. Glymour, Spirtes, and Richardson argue, not without some merit, that in this setting epidemiologists may be unwilling to conclude that X and Y are truly uncorrelated because the power of the tests may be poor due to various forms of measurement error, i.e., outliers, rounding error, etc. (Since many analysts know to use robust nonparametric tests—such as the Wilcoxon test—we do not believe minor violations of distributional assumptions should be the issue.)

Suppose Glymour, Spirtes, and Richardson are correct here. Then, on the one hand, this poor power will, as we have argued, lead to incorrect inferences based on a faithfulness analysis as implemented in Tetrad. On the other hand and even more importantly, as discussed just above and in the second paragraph of our section 7, the main reason that epidemiologists believe that essentially any two variables will be correlated is because of the highly statistically significant associations that are the rule in large observational studies. Indeed, the ubiquity of such statistically significant empirical associations even in the presence of the random (nondifferential) measurement error argued by Glymour, Spirtes, and Richardson is strong evidence for, not against, the proposition that essentially any two variables are correlated.

We now consider the question whether after reading Glymour, Spirtes, and Richardson's reply, practicing epidemiologists should continue to believe that essentially any two variables are correlated due to residual unmeasured

confounders. We believe the answer is yes. However, in their section 2, Glymour, Spirtes, and Richardson make a philosophical argument against the following philosophical claim which we make in the second paragraph of our section 7: Given any two variables X and Y , the universe contains so many unmeasured potential common causes that it is a priori highly unlikely that not a single one is an actual common cause. We were not persuaded by their philosophical argument against this claim. More importantly, however, even had we been sympathetic to their argument, we would reject its conclusion when confronted with the empirical evidence that in most large observational studies, most variables are highly statistically significantly associated.

4. Conclusions Re: Detection of Small Effects

In our chapter, we demonstrated that if the prior probability is near one that any two variables are correlated due to unmeasured confounding, then small causal effects can neither be ruled in nor ruled out based on observational data. In their concluding paragraph, Glymour, Spirtes, and Richardson agree that small effects cannot be either ruled in or out (although, as discussed in detail above, they apparently disagree with the logic by which we reach this conclusion). Glymour, Spirtes, and Richardson go on to conclude that it would be useful to have studies which examine the sensitivity of the output of Tetrad to different assumptions and priors, using as a measure of success the difference between the predicted and actual causal influences. We agree that such studies are necessary to begin to validate the usefulness of Tetrad, since, up to the present, the only justification offered for reliance on Tetrad is asymptotic under the disputed assumption that the probability of there being no unmeasured confounders is not small in relation to sample size. As evidence of our enthusiasm for sensitivity studies, we are collaborating with Glymour, Spirtes, and Richardson and their coworkers in the design of these studies. However, until such sensitivity studies have been conducted, we believe that inferences made by the Tetrad program should be viewed with great skepticism.

5. Comparison of Tetrad and Standard Adjustment Re: Detection of Large Effects

In this section, we compare, by means of examples, standard adjustment with Tetrad as tools for detecting moderate to large causal effects. We believe that

familiarity with these examples will help a user to critically interpret the output of Tetrad. The following two paradigmatic examples make us leery about the use of Tetrad to help detect even large causal effects. As our first paradigmatic example, consider a nonexperimental study of the effect of a new surgical procedure Y on a life-threatening illness Z . Often associations (even adjusted for measured variables) will be large in such studies (especially when they are conducted by the surgeon who developed and believes in the new procedure), even though later randomized trials prove that the surgical procedure has no beneficial effect. The reason for this is that surgeons tend to operate on healthier patients. In a standard (non-Tetrad) analysis, one would adjust for measured presurgical indicators of health status. However, even were a large adjusted apparent beneficial effect for the surgical procedure found in the data, these results would be treated with skepticism since, unlike the cigarette smoking example of our chapter, our prior probability of there being no strong unmeasured confounders would not be small, since surgeons often can recognize good operative candidates based on subtle clues not recorded in the medical chart. Only randomized experiments would be conclusive.

However, suppose now that our surgeon had access to Tetrad and he used the fast causal inference (FCI) algorithm, assuming (correctly) no selection bias and imposing the constraints implied by our knowledge of the temporal ordering of the variables. The FCI algorithm would then search and, if it found some pretreatment variable X such as ethnicity, socioeconomic status, etc., that was correlated with having the operation Y and surgical success Z but not significantly associated with surgical success conditional on Y (as described in our section 5), it would conclude that the operation was causally beneficial. However, while under Spirtes, Glymour, and Scheines's asymptotics this should not occur, we would not be surprised if such occurrences were common due to poor power and the large number of covariates X . The surgeon would then have Tetrad's support for his treatment being beneficial and might then argue that conducting the randomized trial (which would establish the uselessness of his treatment) was now unethical. It is precisely because, in contrast to standard adjustment methods, Tetrad appears (we believe inappropriately) to be able to rule out the hypothesis of no unmeasured confounders in examples like these, that we are wary about its use.

Note in contrast to Glymour, Spirtes, and Richardson's implication in their reply, the standard adjustment approach and the Tetrad approach are not roughly comparable when the goal is to detect rather large causal effects. In fact, the basic semantics of the two approaches differ. Specifically, the standard adjustment point of view simply says that if there are no strong unmeasured confounders, then strong associations are causal; it does not presume to infer from the data that there are indeed no strong unmeasured confounders. The standard approach can be usefully generalized by incorporating sensitiv-

ity analyses that indicate how one's inferences should change under different assumptions about the magnitude of unmeasured confounding (Robins 1997), still, without attempting to infer the true magnitude of residual confounders from the data. Analyses of nonexperimental data with Tetrad to determine the frequency of disagreement with subsequent randomized trial data would be an important test of the Tetrad program and one we understand Glymour, Spirtes, and Richardson and coworkers hope to conduct.

As our second paradigmatic example, consider now the obverse of the previous example. Suppose now Y is cigarette smoking, Z is lung cancer, and data are again available on many variables X that precede Y and Z such as ethnicity or socioeconomic status. Suppose a cigarette company hopes to show that cigarette smoking Y does not cause lung cancer Z despite their very large and striking association. To do so, they use the FCI algorithm in Tetrad to discover a variable X which is associated with lung cancer Z conditional on Y but is marginally independent of Z . Then, by faithfulness and the fact that Y is prior to Z , the Tetrad program will report that it has discovered (1) that Y is not a cause of Z and (2) that the strong empirical association between Y and Z was entirely attributable to one or more strong unmeasured common causes. Now if in fact Y does cause Z , Spirtes, Glymour, and Scheines's asymptotics imply that the above scenario should not occur. However, again we would not be surprised if it did occur due to poor power and the large number of potential covariates X . Again, in this example, the standard adjustment approach and the Tetrad approach are not roughly comparable.

Now one response that an advocate of Tetrad might have to the above examples is that they are unsurprising. For, after all, Tetrad, in contrast to a standard analysis, is trying to do more – it is trying to discover causal structure. Hence, since it risks taking a position on causal structure, it, of course, risks being wrong (particularly in the face of a multiple testing problem presented by having many covariates X). We agree there can be no “risk free” way to search for causal structure. However, since we have no way in general to validate whether taking such risks leads to benefit or harm on average, it seems inappropriate to advocate for Tetrad.

In the next two paragraphs, we shall show that, when our goal is to detect large causal effects when there are no strong unmeasured confounders, there are other paradigmatic settings in which the two approaches are comparable. However, they are often comparable in the sense that Tetrad reduces to and thus adds nothing to the standard adjustment approach.

As our next paradigmatic example, again suppose that X is temporally prior to Y and the magnitude of the $X - Y$ association in the data is small or nonexistent (irrespective of whether the p -value for a test of independence between X and Y is or is not extreme). Then, from the standard adjustment point of view, if we are essentially certain there are no strong unmeasured confounders, we conclude that the magnitude of the causal effect of X and Y

must be small, regardless of whether a test for the independence of X and Y does or does not reject. This reflects the fact that if the causal effect were large, then the lack of a strong $X - Y$ association would require a strong unmeasured confounder. It follows in this setting that it is irrelevant to test for independence between X and Y . However, such a test is the sine qua non of the Tetrad procedure. That is, it is what defines the Tetrad procedure as possibly different from the standard approach.

As our next paradigmatic example, consider again the setting of section 5 where X , Y , and Z are temporally ordered, all correlated, and the magnitude of the $Y - Z$ association controlling for X is small. Then, under the assumption of no strong unmeasured confounders, the standard adjustment approach would declare that the magnitude of the causal effect of Y on Z is small, regardless of whether a statistical test for the independence of X and Z given Y does or does not reject. Thus, performing such a test (which is the sine qua non of a Tetrad analysis) adds nothing to our causal conclusions. In this same setting, suppose now that the conditional association between Y and Z given X were large. Then a standard adjustment approach, under the assumption of no strong unmeasured confounders, would conclude that Y is an important cause of Z , irrespective of whether or not a test for conditional independence between X and Z given Y rejects. Thus again, Tetrad has nothing additional to offer.

As an example where Glymour, Spirtes, and Richardson might still argue that Tetrad might be useful even under the assumption of no strong unmeasured confounders, consider the setting described in the last paragraph but now assume that $X = (X_1, \dots, X_n)$ is highly multivariate with n components. To further simplify the problem, suppose that we know that there are no unmeasured confounders. Since X is highly multivariate, standard regression techniques are unsuitable because of huge standard errors. In this setting, Tetrad could be used to try to determine that certain subsets of the covariates in X are not confounders and thus can be eliminated by conducting preliminary tests of (1) unconditional independence between various components of X and/or (2) conditional independence between components of X and Z given Y . However, a preliminary test approach to the control of confounding has a long history in statistics and is widely viewed as an inadequate approach, as it does not approximate a Bayes approach except for certain specific priors. Further, even for those priors, it is not Bayes, since preliminary test estimators are not admissible. These issues are well discussed for the normal model by Leamer (1978). Thus, it is not clear that Tetrad will be useful for model selection, even in a setting in which there are no unmeasured confounders and we are interested in large causal effects.

In summary, as we have urged, Glymour, Spirtes, and Richardson have abandoned justifying the use of Tetrad based on an asymptotics that assumes the probability of no unmeasured confounders is not small in comparison to

sample size and have accepted our argument that small causal effects cannot be ruled in or ruled out from observational data. Given these admissions, it seems that a great deal of work will be necessary to justify, if possible, the usefulness of Tetrad as a reliable search engine for causal effects. We would guess that if such a justification can be formulated, it will be in association with a highly modified and more cautious Tetrad algorithm.

References

- Hill, A. B. 1951. *Principles of Medical Statistics, 9th Edition*. New York: Oxford University Press.
- Leamer, E. 1978, *Specification Searches*. New York: John Wiley & Sons.
- Robins, J. M. 1997. Causal Inference from Complex Longitudinal Data. In *Latent Variable Modeling and Applications to Causality. Lecture Notes in Statistics (120)*, ed. M. Berkane, 69-117. Berlin: Springer-Verlag.
- Robins, J. M., Landrigan P. J., Robins T. G., and Fine L. J. 1985. Decision-Making under Uncertainty in the Setting of Environmental Health Regulations. *Journal of Public Health Policy*, 6(3): 322-328.
- Spirtes, P.; Glymour, C.; and Scheines, R. 1993. *Causation, Prediction, and Search*. New York: Springer-Verlag.